XIX Jornadas de Classificação e Análise de Dados
Instituto Politécnico de Tomar, 28 – 31.Março.2012

Não preencher

Avaliação do Desempenho em Análise Discriminante Discreta

Anabela Marques[1], Ana Sousa Ferreira[2], Margarida Cardoso[3]
[1]*Escola Superior de Tecnologia do Barreiro, Instituto Politécnico de Setúbal, e-mail: anabela.marques@estbarreiro.ips.pt;*
[2]*LEAD, Faculdade de Psicologia, Universidade de Lisboa, UNIDE e CEAUL, e-mail: asferreira@fp.ul.pt;;*
[3]*Dep. de Métodos Quantitativos do ISCTE- Instituto Universitário de Lisboa, UNIDE, e-mail: margarida.cardoso@iscte.pt*

Sumário

A Análise Discriminante Discreta (ADD) está frequentemente associada a estudos nas áreas das ciências sociais e da saúde. Nestes domínios é comum dispor de classes *a priori* mal separadas e/ou amostras de pequenas dimensões. Neste contexto, muitos dos métodos de ADD revelam um fraco desempenho, impondo-se o desenvolvimento de outros métodos de classificação, nomeadamente por recurso à combinação de modelos.

Neste trabalho iremos avaliar o desempenho de um método de ADD, usando uma abordagem de combinação de modelos, recorrendo à taxa de observações corretamente classificadas e a uma medida de associação entre as classes *a priori* e as classes previstas segundo a análise efetuada. Estas medidas serão determinadas em amostra de teste e/ou mediante validação cruzada.

Palavras-chave Análise Discriminante Discreta; Combinação de modelos; Modelo de Emparelhamento Hierárquico; Modelo Gráfico Decomponível; Modelo de Independência Condicional;

1. Introdução

A Análise Discriminante Discreta (ADD) é uma técnica de análise de dados multivariados que se aplica a populações munidas de uma partição definida *a priori*, sendo cada elemento da população descrito por um conjunto de variáveis qualitativas, e tem por objetivo quer conhecer as variáveis que melhor discriminam as classes da partição, quer a definição de uma regra de decisão que permita classificar correctamente novos elementos.

Em ADD, o modelo mais natural é o Modelo Multinomial Completo (MMC) (Goldstein e Dillon, 1978) que assume que as funções de probabilidade em cada classe são leis de

probabilidade multinomiais. Neste caso, as funções de probabilidade condicionais são estimadas pelas frequências observadas em cada classe.

Embora este modelo seja naturalmente adequado aos dados observados em ADD, exige contudo a estimação de um grande nº de parâmetros. Por exemplo, para seis variáveis binárias em estudo teremos 2^6 -1 = 63 parâmetros a estimar, em cada classe. Logo, mesmo para um número reduzido de variáveis em estudo a estimação fiável dos parâmetros deste modelo obriga a amostras de grandes dimensões.

Uma das abordagens para limitar o problema da dimensionalidade no modelo MMC consiste na redução da complexidade do modelo. No Modelo de Independência Condicional de Ordem Um (MIC), assume-se que as variáveis em análise são independentes dentro de cada classe. Desse modo, o número de parâmetros a estimar reduz-se drasticamente conduzindo apenas, no nosso exemplo das seis variáveis binárias, à estimação de seis parâmetros em cada classe.

Este modelo de classificação revela, geralmente, um bom desempenho embora seja irrealista considerar, em muitos estudos, que as variáveis descritoras são independentes dentro de cada classe.

Os modelos MMC e MIC são considerados por muitos autores (Goldstein e Dillon, 1978; Celeux e Nakache, 1994) modelos de referência em ADD uma vez que se situam em pólos opostos quer quanto às hipóteses que assumem sobre as relações entre as variáveis em estudo quer quanto ao número de parâmetros a estimar. Ao longo das últimas décadas, diversos modelos de classificação foram propostos, inspirados nos modelos MMC e MIC.

Um desses modelos é o Modelo Gráfico Decomponível (MGD) (Celeux e Nakache, 1994; Pearl, 1988) que, tal como o modelo MMC considera a relação entre as variáveis em estudo, baseando-se num conceito de árvore de dependência (Chow e Liu, 1968).

2. Avaliação de uma combinação de modelos em Análise Discriminante Discreta

A combinação de modelos surgiu por volta da década de 90 com o objectivo de encontrar métodos que se adaptassem melhor ao comportamento dos dados em estudo e que pudessem minimizar o número de parâmetros a estimar. O interesse nesta abordagem tem vindo a crescer em Análise Discriminante, nomeadamente nas áreas das ciências sociais e da saúde, onde a recolha de amostras de dimensão razoável é por vezes inexequível de ser atingida. Os resultados obtidos por Sousa Ferreira, (Sousa Ferreira, 2000; Sousa Ferreira *et al*., 2000), nos trabalhos desenvolvidos em Análise Discriminante Discreta mostraram que a abordagem pela combinação de modelos conduzia a modelos mais eficientes e estáveis, tanto mais que frequentemente se observava que os erros de afetação obtidos por vários modelos não ocorriam sobre os mesmos objetos (Brito *et al*., 2006).

Marques *et al*. (2008) propuseram uma combinação linear entre o Modelo de Independência Condicional (MIC) e o Modelo Gráfico Decomponível (MGD), recorrendo a um único coeficiente β ($0 \leq \beta \leq 1$), conduzindo a um modelo intermédio entre estes dois modelos de classificação.

Neste trabalho pretende-se avaliar o desempenho da combinação de modelos proposta por Marques *et al*. (2008).

A avaliação do desempenho em ADD baseia-se, habitualmente, na taxa das observações corretamente classificadas estimada quer na amostra-base (estimação por resubtituição), quer em amostra-teste ou ainda por validação cruzada.

Diversos autores (Paik, 1998; Sousa Ferreira e Cardoso, 2009) têm vindo a propor novas medidas de avaliação do desempenho em ADD, procurando usar toda a informação disponível na matriz de confusão onde habitualmente se registam os resultados obtidos por um modelo de ADD e não apenas a sua diagonal principal, onde se observam o nº de observações corretamente classificadas

Neste trabalho procura-se comparar a avaliação do desempenho da combinação de modelos MIC-MGD, recorrendo quer à taxa de observações corretamente classificadas quer a uma medida de associação entre as classes *a priori* e as classes previstas segundo a análise efetuada (Paik, 1998).

Para avaliar o desempenho do referido modelo, recorremos quer a dados reais quer a dados simulados, considerando em ambos os casos seis variáveis binárias.

Os dados reais considerados consistem numa amostra de 34 pacientes do foro dermatológico, classificados em três classes segundo o seu grau de alexitimia: Não Alexítimicos (C_1), Alexitímicos (C_2) e Intermédios (C_3). Esta classificação nas três classes baseou-se na Escala de Alexitimia de Toronto (TAS-20) e procurou-se compreender como é que a Alexitimia[1] se traduzia no Teste do Rorschach, um teste psicológico projetivo de personalidade.

Os dados simulados foram obtidos com base no modelo de Bahadur (Goldstein e Dillon, 1978), considerando amostras de pequena e moderada dimensão, 120 e 400 observações respetivamente. Estas amostras satisfazem dois tipos de estrutura: independência – IND – gerada a partir do modelo MIC e de correlação – DIF – gerada considerando a existência de diferentes relações entre as variáveis nas várias classes. Para ambos os casos considerámos duas e quatro classes definidas *a priori*.

[1] Alexitimia consiste na dificuldade para expressar e descrever emoções.

Referências:

BACELAR-NICOLAU, H. (1985) The Affinity Coefficient in Cluster Analysis. *Meth.Oper. Res*., 53, 507-512.

BRITO, I., CELEUX, G. & SOUSA FERREIRA, A. (2006) Combining method in supervised classification: A comparative study on discrete and continuous problems. *REVSTAT - Statistical Journal,* Vol. 4(3), 201-225.

CELEUX, G. and NAKACHE, J.P. (1994) *Analyse Discriminante sur Variables Qualitatives*. G. Celeux et J. P. Nakache (Eds.), Polytechnica.

GOLDSTEIN, M., DILLON, W.R. (1978) *Discrete Discriminant Analysis*, Wiley and Sons.

MARQUES, A., SOUSA FERREIRA, A. & CARDOSO, M. (2008) Uma proposta de combinação de modelos em Análise Discriminante. IN OLIVEIRA, I. et al. (Eds.) *"Estatística - Arte de Explicar o Acaso", Ciência Estatística*, Edições S.P.E., 393-403.

MATUSITA, K. (1955) Decision Rules Based on Distance for Problems of Fit, Two Samples and Estimation. *Ann. Inst. Stat. Math*.,. 26(4), 631-640.

PAIK, H. (1998) *The Effect of Prior Probability on Skill in Two-Group Discriminant Analysis*. In Quality and Quantity, 32, 201-211.

PEARL, J. (1988) *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Los Altos: Morgan Kaufmann.

PRAZERES, N.L. (1996) *Ensaio de um Estudo sobre Alexitimia com o Rorschach e a Escala de Alexitimia deToronto (TAS-20)*. Master Thesis, Univ. Lisbon.

SOUSA FERREIRA, A. (2010) A Comparative Study on Discrete Discriminant Analysis through a Hierarchical Coupling Approach. IN LOCAREK-JUNGE, H., WEIHS, C. (Eds.) *Classification as a Tool for Research,* Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag, Heidelberg-Berlin, 137-145.

SOUSA FERREIRA, A. (2000) *Combinação de Modelos em Análise Discriminante sobre Variáveis Qualitativas*, Tese de doutoramento, Univ. Nova de Lisboa.